Available online at www.sciencedirect.com

SciVerse ScienceDirect

Electronic Notes in Theoretical Computer Science

Electronic Notes in Theoretical Computer Science 275 (2011) 23–36

www.elsevier.com/locate/entcs

The Monte Carlo EM method for the parameter estimation of biological models

Alessio Angius[1] András Horváth[2]

Department of Computer Science, University of Torino, Torino, Italy

Abstract

It is often the case in modeling biological phenomena that the structure and the effect of the involved interactions are known but the rates of the interactions are neither known nor can easily be determined by experiments. This paper deals with the estimation of the rate parameters of reaction networks in a general and abstract context. In particular, we consider the case in which the phenomenon under study is stochastic and a continuous-time Markov chain (CTMC) is appropriate for its modeling. Further, we assume that the evolution of the system under study cannot be observed continuously but only at discrete sampling points between which a large amount of reactions can occur.
The parameter estimation of stochastic reaction networks is often performed by applying the principle of maximum likelihood. In this paper we describe how the Expectation-Maximisation (EM) method, which is a technique for maximum likelihood estimation in case of incomplete data, can be adopted to estimate kinetic rates of reaction networks. In particular, because of the huge state space of the underlying CTMC, it is convenient to use such a variant of the EM approach, namely the Monte Carlo EM (MCEM) method, which makes use of simulation for the analysis of the model. We show that in case of mass action kinetics the application of the MCEM method results in an efficient and surprisingly simple estimation procedure. We provide examples to illustrate the characteristics of the approach and show that it is applicable in case of systems of reactions involving several species.

Keywords: parameter estimation, mass action kinetics, maximum likelihood, expectation-maximisation method

1 Background

As described by Gillespie in [5] the temporal behaviour of a biochemical system can be described by a stochastic process, in particular, by a continuous time Markov chain (CTMC). In order to have a complete description of the CTMC model describing the phenomenon under study and to be able to perform its analysis, the estimation of the kinetic rates is a *"conditio sine qua non"*. In this context, the parameter estimation is essentially an optimisation problem which aims to find the set of parameter such that the model is able to reproduce the experimental observations with high probability. The problem is not trivial for several reasons. The

[1] Email: angius@di.unito.it

[2] Email: horvath@di.unito.it

doi:10.1016/j.entcs.2011.09.003

studied phenomenon can be very complex with several reagents interacting through many reactions. Moreover, it is often unrealistic to consider the process as perfectly and continuously observable. In particular, the measurement techniques are often unable to observe the system behaviour as a continuous process and provide observations of the system state only at a limited set of time instants. Moreover, consecutive time instants can be so far from each other that a considerable amount of reactions occur between them. This means that we have to face an optimisation problem with incomplete data in hand. A method to maximum likelihood estimation in case of incomplete data, namely the Expectation-Maximisation (EM) method, has been presented by A. Dempster in [3]. The basic idea of EM method is to rebuild the missing data in expectation and apply optimisation to find parameters that maximises the probability of the reconstructed complete data. It is often the case that the exact reconstruction of the missing data is a hard task. In these cases, as proposed by Wei and Tanner in [15], simulation can be used to complete the data and this approach is called the Monte Carlo EM (MCEM) method.

In this work we adopt the MCEM method to the estimation of kinetic rates in stochastic reaction networks. In particular, we consider stochastic reaction networks evolving according to mass action kinetics and show that the MCEM method leads to a simple and efficient estimation procedure.

Several works exist on estimating kinetic rates of reaction networks by applying optimisation methods [10,4,12]. Most of these works however do not consider stochasticity but apply a deterministic view of the evolution of the phenomenon under study. In theory, it is possible to transform the rates obtained for the deterministic model into rates that can be used in a stochastic setting but, as it is pointed out in [11], this is not always possible. This observation led to attempts to give an estimate of the kinetic rates in accordance with the stochastic view introduced by [5]. Bayesian inference methods were used in [7,1], maximum likelihood methods were applied in [11,2,13]. The strength of our approach, with respect to the ones cited above, is that it works with limited information (i.e., it is possible to apply it with very infrequent observations between which thousands of reactions occur) and does not involve heavy optimisation tasks.

The paper is organised as follows. In the next section we provide the reference stochastic model. Then a brief introduction of the EM method in general is given. Subsequently, we describe the application of the MCEM method to the estimation of parameters of stochastic reaction networks. The last but one section is dedicated to the numerical illustration of the proposed approach. In the last section we draw the conclusions.

2 Considered model

We consider a continuous time Markov chain (CTMC) describing the interaction of M reagents through R reactions. The state of the system is given by a vector of M integers providing the quantity of the reagents. The effect of reaction $i, 1 \leq i \leq R$ is described by a vector of M integers denoted by $\overline{e}_i$ in such a way that if reaction

i occurs in state $\overline{x}$ then the next state is $\overline{x}' = \overline{x} + \overline{e}_i$. The transitions of the CTMC represent the reactions and their intensity depends on the state of the system. By $f_i(\overline{x})$ we denote the function providing the intensity of reaction i in state $\overline{x}$. Throughout the paper we assume that the intensities are of the form

$$f_i(\overline{x}) = k_i \prod_{j=1}^{M} \binom{x_j}{a_{i,j}} \tag{1}$$

where the constants $a_{i,j}$ provide the stoichiometry of the ith reaction with $a_{i,j} \in \mathbb{N}$ and k_i is the kinetic rate constant (i.e., it describes the speed of the interaction). Consequently, the model corresponds to mass action kinetics and the CTMC is exactly the one simulated by the classical algorithm of Gillespie [5].

As an example of the model consider the following system of reactions [6] which will be used among the numerical examples as well:

$$\begin{aligned} DNA_{off} &\xrightarrow{k_1} DNA_{on}, \\ DNA_{on} &\xrightarrow{k_2} DNA_{off}, \\ DNA_{on} &\xrightarrow{k_3} mRNA + DNA_{on}. \end{aligned}$$

The above set of reactions describes that DNA is switched on/off by polymerase binding/unbinding and polymerase bound (i.e., switched-on) DNA is transcribed into $mRNA$. Switch on is described by the first reaction transforming DNA_{off} to DNA_{on}, switch off is described by the second reaction transforming DNA_{off} to DNA_{on} and transcription is due to the third reaction which produces $mRNA$ leaving the actual quantity of DNA_{on} unchanged. As three reagents are involved, the state of the system is a triple $\overline{x} = |x_1, x_2, x_3|$ describing the quantities of DNA_{off}, DNA_{on} and $mRNA$, respectively. The vectors describing the effect of the three reactions are: $\overline{e}_1 = |-1, 1, 0|$, $\overline{e}_2 = |1, -1, 0|$ and $\overline{e}_3 = |0, 0, 1|$. The intensities associated with the reactions assuming mass action kinetics are: $f_1(\overline{x}) = k_1 x_1$, $f_2(\overline{x}) = k_2 x_2$ and $f_3(\overline{x}) = k_3 x_2$. Based on the above description, the stochastic simulation of the system is a straightforward task.

3 EM method

The Expectation-Maximisation (EM) method is an algorithm for maximum likelihood parameter estimation in case of incomplete data. The input of the EM algorithm is composed by a set of samples and a stochastic model characterised by a set of parameters denoted by λ.

The EM method is iterative, i.e., it starts from an initial guess of the set of parameters, λ_0, and then improves it step by step in such a way that the behaviour provided by the model is more and more similar to the behaviour described by the samples. The set of parameters after i steps is denoted by λ_i. Each iteration is composed of two steps called, respectively, Expectation step (E-step) and Maximisation step (M-step).

The role of the E-step is to compute the missing information in expectation. Formally, denoting by Y the set of incomplete data, and by Z the complete data, the E-step computes the conditional expectation $E[Z|Y, \lambda_i]$. In the context of our problem, Y contains the samples at discrete time points and the E-step aims to calculate, given the current set of parameters in λ_i, the most typical full trajectory, Z, that goes through the observed samples given in Y.

The M-step is applied then to find a new set of parameters λ_{i+1} such that the likelihood of the trajectory generated during the E-step is maximal. Once the new set of parameters λ_{i+1} is found, it is used as the starting point for the next iteration.

In many situations, including the one considered in this paper, the strength of the EM method lies in the fact that finding such parameters that maximise the likelihood of the incomplete data is much harder than finding parameters that maximise the likelihood of the complete data. In other words, the optimisation required in the M-step is less burdensome than the original optimisation problem. In particular, in relation to the problem considered in this paper, the M-step is very simple in the case of mass action kinetics. In turn, in many cases, including ours, the E-step can be hard both from a theoretical and computational point of view.

For those cases in which the computations required by the E-step are particularly complex, a variant of the EM method can be applied. In this variant, the exact computation of the conditional expectation in the E-step is substituted by simulation. This approach is known as the Monte Carlo EM (MCEM) method and it is particularly useful in situations when performing the E-step in an exact manner is either too time consuming or even unfeasible. For the problem considered in this paper, because of the huge state space of the involved CTMC, the only viable approach is provided by the MCEM method.

The convergence characteristics of the MCEM method are poorer than those of the EM method and the simulation can introduce fluctuations of the parameters. It is still possible however to prove that the convergence of the method is preserved if the number of iterations is high [15].

4 MCEM for biochemical systems

Problem formulation

We assume that we are given a network of reactions and experimental observations of quantities of the involved species at discrete time instants. The time instants of the observations will be denoted by $t_0 = 0, t_1, t_2, \ldots, t_N$ and the associated observations by $\overline{y}_0, \overline{y}_1, \ldots, \overline{y}_N$ where $\overline{y}_i$ is a vector of integers providing the state of the system at t_i. (We consider here only a single sequence of observations but the extension to multiple observation sequences is straightforward.) We assume that all or some of the kinetic rates are not known, i.e., there are unknown constants in the functions $f_i(\overline{x})$ which provide the intensity of the reactions. The set of these unknown constants will be denoted by λ and we will write $f_i(\overline{x}, \lambda)$ to make explicit the dependence of the intensities on the unknowns. Our aim is to give a maximum likelihood estimate for the unknown kinetic rates by the MCEM method. In the

following two subsections we describe the E-step and the M-step.

E-step

Given the set of samples and the current estimate of the parameters, the E-step aims to build the most probable full trajectory that goes through the observed states. This requires to find the most probable trajectory between each two consecutive sample points. The E-step hence requires to find most likely random walks over CTMCs. This is possible in theory but, unfortunately, as in our context the considered CTMC almost always has a huge state space, it cannot be performed in an exact manner. As anticipated, in this situation the E-step can be solved by simulation which provides a good approximation of the most probable trajectory. Generating trajectories of a CTMC by simulation is straightforward.

As we are given $N+1$ samples, we need N subtraces to connect the observation time instants. As the CTMC is huge and the current set of parameters can be far from the real set of parameters, it is very unlikely that a single simulation run arrives exactly (or even close) to the observed states. For this reason the E-step is composed by the following two phases.

(i) **Generation of traces.** For each interval $[t_i, t_{i+1}], 0 \leq i \leq N-1$, we generate K random walks of length $t_{i+1} - t_i$ starting from $\overline{y}_i$ and choose the one that arrives closest in "distance" to $\overline{y}_{i+1}$. The concept of "distance" between the sample point and the last state of the random walk is expressed as the sum of the relative errors over the species.

(ii) **Improvement of traces.** In this phase we improve the subtraces by modifying them. We pick up randomly a reaction from the subtrace and check if it can be substituted by another reaction in such way that the subtrace arrives closer to the observed state. The substitution is accepted only if all the remaining reactions are still possible. The times between consecutive reactions remain unchanged. The extent of the modification is determined by a parameter $\rho \in [0, 1]$ which defines the proportion of the reactions that we attempt to substitute.

The proportion defined by ρ has a delicate role in the estimation process. When the current estimate is far from the real parameters, a higher ρ is necessary in order to come up with reasonable subtraces and to have faster convergence of the estimation process. Instead, when the estimates are already good, a lower ρ (even $\rho = 0$) has to be used in order to not to alter too much the stochastic behaviour induced by the actual estimate.

In terms of complexity, the cost of the generation of random walks is linear in the number of reactions occurrences, and the storage of the best random walk is very cheap, since each subtrace can be uniquely identified by means of the seed of the pseudo-random number generator. For these reasons, the improve of the best trace represents the most expensive phase of the method, since each trial forces the

unroll and the check of the enabling of all subsequent reactions [3] . As a consequence, the implementation of this phase could be not trivial. A first hint could be the use of a preprocessing which after the random selection of the "candidates" for the substitution sorts them in order of occurrence. In this manner the whole trace can be unrolled just ones and for each attempt the number of checks becomes smaller and smaller. The second hint is to verify more than one reaction at a time and discard (without additional costs) all substitutions if the control is not satisfied. This solution could discard some valid substitutions but is more convenient in terms of computation time.

M-Step

In the M-step we have to find the next set of estimates, λ_{i+1}, that maximises the likelihood of the full trajectory generated in the E-step. The ith subtrace generated by the E-step, reconstructing the most probable trajectory between state $\overline{y}_{i-1}$ and state $\overline{y}_i$ in the time interval $[t_{i-1}, t_i]$, will be denoted by S_i and it has the form

$$\overline{y}_{i-1} = \overline{s}_{i,1} \overset{r_{i,1},u_{i,1}}{\longrightarrow} \overline{s}_{i,2} \overset{r_{i,2},u_{i,2}}{\longrightarrow} \cdots \longrightarrow \overline{s}_{i,H_i} \overset{u_{i,H_i}}{\longrightarrow} \overline{s}_{i,H_i}$$

where H_i denotes the length of the ith subtrace and $s_{i,j}, r_{i,j}$ and $u_{i,j}$ are the states, the reactions and the sojourn times of the ith subtrace, respectively. We have that

$$\sum_{j=1}^{H_i} u_{i,j} = t_i - t_{i-1}.$$

The last arrow is without a reaction and this represents the fact that the process remains in state $\overline{s}_{i,H_i}$ for at least u_{i,H_i} time units. The closer state $\overline{s}_{i,H_i}$ is to state $\overline{y}_i$ the better the ith subtrace reflects the observed behaviour.

It follows from the theory of CTMCs that the likelihood of the ith subtrace, denoted by L_i, can be calculated as

$$L_i = \left(\prod_{j=1}^{H_i-1} f_{r_{i,j}}(\overline{s}_{i,j}, \lambda) e^{-f(\overline{s}_{i,j},\lambda)u_{i,j}} \right) \cdot e^{-f(\overline{s}_{i,H_i},\lambda)u_{i,H_i}} \tag{2}$$

where

$$f(\overline{x}, \lambda) = \sum_{k=1}^{R} f_k(\overline{x}, \lambda)$$

is the sum of the intensities of the reactions in state $\overline{x}$. The product in (2) gives the likelihood of the transitions of the ith subtrace while the last exponential term is the probability that the process does not leave state $\overline{s}_{i,H_i}$ for at least u_{i,H_i} time

[3] Note that the unroll of the trace is expensive almost as its generation.

units. The likelihood of all the subtraces is simply given by

$$L = \prod_{i=1}^{N} L_i$$

and we have to find such λ that maximises this product.

In order to find the maximum of L, it is useful to take its logarithm in which products are transformed into sums as

$$\ln(L) = \sum_{i=1}^{N} \ln(L_i) = \sum_{i=1}^{N} \left(\sum_{j=1}^{H_i - 1} \ln(f_{r_{i,j}}(\overline{s}_{i,j}, \lambda)) - \sum_{j=1}^{H_i} f(\overline{s}_{i,j}, \lambda) u_{i,j} \right). \quad (3)$$

Naturally, the difficulty of finding the maximum of (3) depends on the functions $f_i(\overline{x}, \lambda), 1 \leq i \leq R$. As mentioned earlier, we consider the case in which the intensity of the reactions corresponds to mass action kinetics (1). Moreover, we assume that the stoichiometry of the reactions (described by $a_{i,j}, 1 \leq i \leq R, 1 \leq j \leq M$) is known which is the typical case in parameter estimation problems. Accordingly, the parameters to estimate are the kinetic rate constants, i.e., $\lambda = \{k_1, ..., k_R\}$. Without loss of generality, we focus our attention on finding such k_1 that maximises (3). Applying (1), the derivative of (3) with respect to k_1 is

$$\frac{\partial \ln(L)}{\partial k_1} = \sum_{i=1}^{N} \left(\sum_{j=1}^{H_i - 1} I\{r_{i,j} \text{ is reaction 1}\} \frac{\prod_{k=1}^{M} \binom{x_k}{a_{1,k}}}{k_1 \prod_{k=1}^{M} \binom{x_k}{a_{1,k}}} - \sum_{j=1}^{H_i} \prod_{k=1}^{M} \binom{x_k}{a_{1,k}} u_{i,j} \right)$$

where I is 1 if its argument is true and 0 otherwise. Denoting by $f_{i,j}$ the number of times reaction j occurs in the ith subtrace we have

$$\frac{\partial \ln(L)}{\partial k_1} = \sum_{i=1}^{N} \left(\frac{f_{i,1}}{k_1} - \sum_{j=1}^{H_i} \prod_{k=1}^{M} \binom{x_k}{a_{1,k}} u_{i,j} \right). \quad (4)$$

It is easy to check that the value of k_1 with which (4) is 0 maximises L. Consequently, the estimate is

$$k_1 = \frac{\sum_{i=1}^{N} f_{i,1}}{\sum_{i=1}^{N} \sum_{j=1}^{H_i} \prod_{k=1}^{M} \binom{x_k}{a_{1,k}} u_{i,j}}. \quad (5)$$

Accordingly, in case of mass action kinetics, the optimisation required in the M-step boils down to the explicit formula given in (5).

Handling other forms of $f_i(\overline{x}, \lambda)$ is out of the scope of this paper. We only mention here that with general $f_i(\overline{x}, \lambda)$ functions the optimisation required by the M-step can become more complex but even in this case it is possible to divide the optimisation problem into smaller subproblems. In order to show this, let us denote by ν_i the variables on which $f_i(\overline{x}, \lambda)$ depends. It is reasonable to assume that the

sets $\nu_i, 1 \leq i \leq R$, are mutually disjoint. In this case the derivative of the log-likelihood function (given in (3)) according to a variable belonging to ν_i does not depend on the variables belonging to the other sets $\nu_j, j \neq i$. This means that the original optimisation problem that involves all the variables can be tackled by R smaller optimisation problems of much smaller dimensions.

5 Illustrative numerical examples

In this section we show numerical results obtained using the MCEM method. We apply the method to two models. For both cases, the samples are generated *"in silico"* by means of simulations. In these tests we put aside the biological meaning of the models and our aim is to illustrate the method and to show that it is able to reconstruct the set of parameters. For all the cases we provide tables to compare the original values and their estimates. Moreover, in order to provide a visual comparison of the behaviours with the original and the estimated values, we provide in figures the evolution of the systems according both to the corresponding ordinary differential equations (ODEs) and to the corresponding stochastic setting. The ODEs are useful to get a quick glimpse of the goodness of the estimates that were obtained in the stochastic setting.

The MCEM method has been implemented in a prototype *JAVA* tool. All the experiments have been performed on a *Intel Centrino Dual Core* with 4Gb of RAM.

5.1 Gene transcription model

Our first example is the model already introduced in Section 2 describing binding and unbinding of the DNA and its transcription into $mRNA$ [6]. We assume to have a single unit of DNA and the initial condition is $[DNA_{off}] = 1$ and $[DNA_{on}] = [mRNA] = 0$.

In order to evaluate the method in different situations, we use the model with different levels of "granularity", i.e., we use different levels of discretizations to obtain discrete models from the originally continuous concentrations. The discretization step will be denoted by h. The initial state of the CTMC modeling the three reactions is $|1/h, 0, 0|$. For the sake of having models that evolve on the same time scale independently of h, the kinetic rate constants have to depend on h. Specifically, the intensity of the reactions are $k_1 h$, $k_2 h$ and $k_3 h$ with $k_1 = 0.027$, $k_2 = 0.166$ and $k_3 = 0.4$. The effect of h is twofold: the smaller h the larger the state space and the less variable the behaviour of the model. Indeed, as h tends to 0, the behaviour of the model tends to the solution of the corresponding ODEs [8,14].

We have generated "in silico" samples in such a way that between consecutive sampling points there are about 25000 reactions. The number of the samples is 25. This means that only a small fraction of what happens in the model is available to the estimation procedure.

The E-step was performed with $K = 20$, i.e., 20 traces were generated in the first phase of the E-step. In the second phase the parameter ρ plays a crucial role. It is convenient to start with a high value of ρ and then to lower it gradually as the

estimates become more reliable. We chose to start with $\rho = 0.4$ and to decrease it as the generated traces get closer to the observations.

Table 1 reports the original parameters and those obtained by the MCEM approach after 100 iterations computed in about a second of CPU time. The initial guess of the parameters was random in the range [0:10]. A possible observation about the results could be that some estimations of k_1 and k_2 are far from the original values. This is caused by the fact that the low number of infrequent discrete samples reflect the ratio $k_1/(k_1+k_2)\times k_3$ (determining the increase of $mRNA$) and, to some extent, the ratio $k_1/(k_1+k_2)$ (determining the quantity of DNA_{on}) but not the value of parameters. More samples with higher sampling frequency could however alleviate this problem. It can also be observed that with finer discretization the results are more accurate. A way of illustrating all the cases on the same figure is to apply the results in the ODEs representing the model. This is depicted in Figure 1. It can be seen that all cases catch well the asymptotic increase rate of $mRNA$ and with $h = 0.001$ the estimates reproduce well the original model. The ODEs could not provide an accurate representation of the CTMC trajectories, for this reason we took in consideration also the stochastic setting. Figures 2 depicts the mean and the variance of the 100.000 simulation traces. As last observation, it is important to point out that by using smaller values of h the method gets better.

	k_1	k_2	k_3
Original	0.027	0.166	0.4
Estimate, h=1	0.6595	1.5386	0.1879
Estimate, h=0.1	0.1279	0.6385	0.3431
Estimate, h=0.01	0.1098	0.6683	0.4115
Estimate, h=0.001	0.0397	0.2591	0.4045

Table 1
Results of the parameter estimation for the gene transcription model

5.2 DFG degradation pathway

In order to test the method with a higher number of variables we use a model which describes the control of the N-(deoxy-D-fructos-1-y1)-glycine (DFG) degradation pathway [9]. The model can be found in the database available on the site www.sbml.org. It involves 14 reagents interacting through 16 reactions with mass action kinetics (the reactions are reported in Table 2). The original ODE model was transformed into a CTMC with discretization step $h = 0.0001$ and started with initial concentration $[DFG] = 9$ and quantity 0 for all other reagents. Note that this choice of h leads to a CTMC with huge state space. The samples contained 20 observations with about 10000 reactions between consecutive samples. The E-step was performed with $K = 20$ and ρ started from 0.2 and decreased throughout the calculations. The results are given in Table 3. It can be seen that the method gives good estimate for almost all the involved parameters. Figure 5 depicts the evolution of some of the reagents of the model according to the corresponding system of ODEs with the original and the estimated parameters. The stochastic setting is depicted in Figure 4 where in honor of synthesis we report the variance only. For

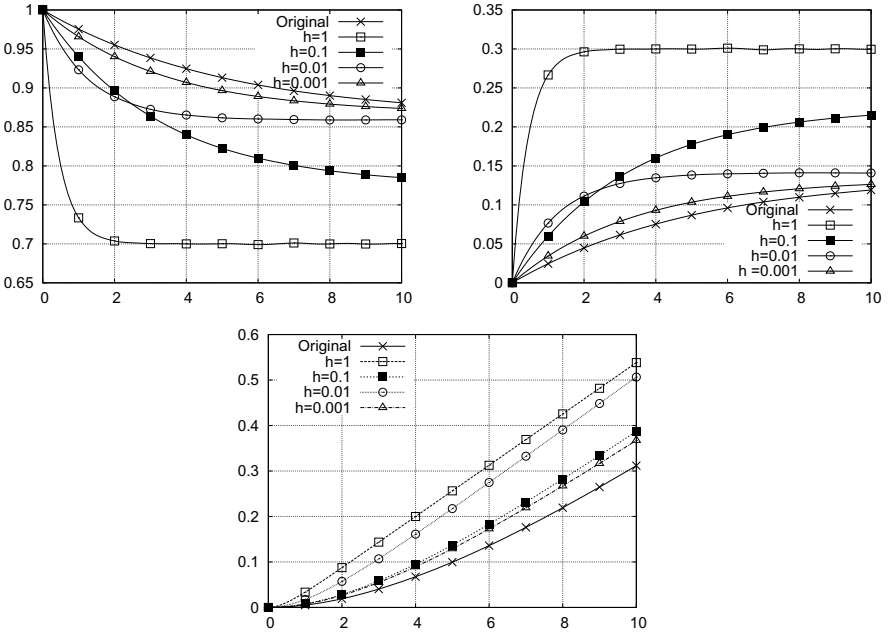

Fig. 1. Gene transcription model: ODE computed with the original parameters and the estimates for different values of h (DNA_{off} on the left, DNA_{on} on the right and $mRNA$ below).

most species there is a good correspondence between the original behaviour and the one provided by the estimate. In Figure 3 we depict instead the likelihood with which the estimates reproduce the original "in silico" generated trace. After 350 interactions the likelihood with the estimates is very close to the likelihood with the original parameters.

Reactions			
$DFG \xrightarrow{k_1} E1$	$DFG \xrightarrow{k_2} E2$	$DFG \xrightarrow{k_3} Gly + Cn$	$E1 \xrightarrow{k_4} Gly + DG3$
$DG3 \xrightarrow{k_5} Cn$	$DG3 \xrightarrow{k_6} wFA$	$E2 \xrightarrow{k_7} Gly + DG1$	$DG1 \xrightarrow{k_8} Cn$
$DG1 \xrightarrow{k_9} AA$	$E1 \xrightarrow{k_{10}} Gly + Man$	$E1 \xrightarrow{k_{11}} Gly + Glu$	$Man \xrightarrow{k_{12}} Glu$
$Glu \xrightarrow{k_{13}} DG3$	$Gly + Cn \xrightarrow{k_{14}} Mel$	$Cn \xrightarrow{k_{15}} AA + FA + MG$	$E2 \xrightarrow{k_{16}} Gly + Fru$

Table 2
Reactions of the DFG degradation pathway

Case 3	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8
Original	0.005	0.015	0.015	0.079	0.090	0.027	0.212	0.181
Result	0.0048	0.0175	0.0119	0.058	0.068	0.010	0.252	0.706
Case	k_9	k_{10}	k_{11}	k_{12}	k_{13}	k_{14}	k_{15}	k_{16}
Original	1.908	0.070	0.113	8.0E-4	0.002	0.003	0.015	0.013
Result	1.847	0.0651	0.122	7.8E-4	0.014	0.003	0.0147	0.0122

Table 3
Result of parameter estimation for the DFG degradation pathway

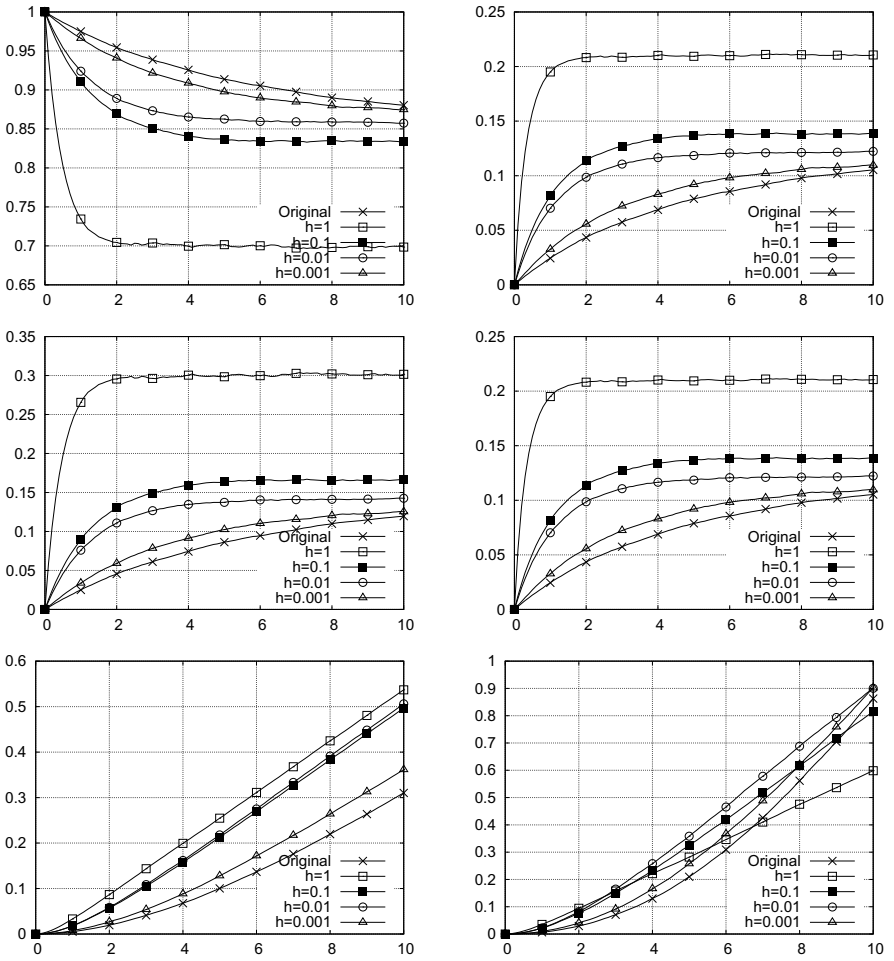

Fig. 2. Gene transcription model: The average (left) and the variance (right) of the quantity of Dna_{off}, Dna_{on}, $mRna$ computed with original parameters and the estimates for different values of h.

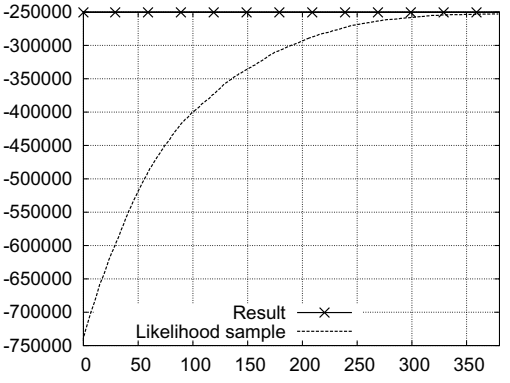

Fig. 3. DFG degradation pathway: likelihood with estimates as function of the number of iterations

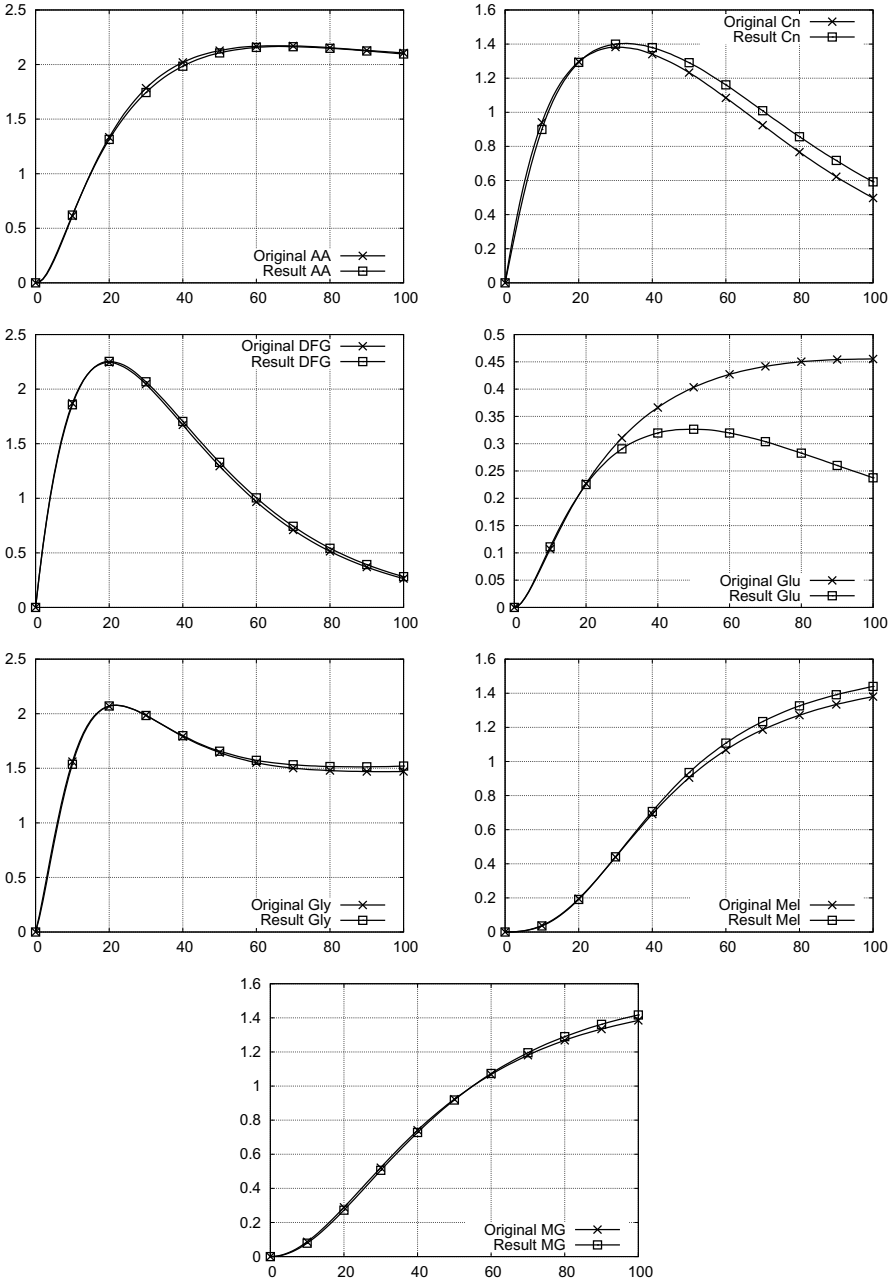

Fig. 4. The variance of some of the reagents involved in the DFG degradation pathway

Conclusions

In this work we adopted the MCEM method to the estimation of kinetic rates in stochastic reaction networks. We have shown that the resulting technique is efficient and leads to surprisingly simple calculations in the case of mass action kinetics. The strength of the proposed approach is that it can be applied even with a limited set

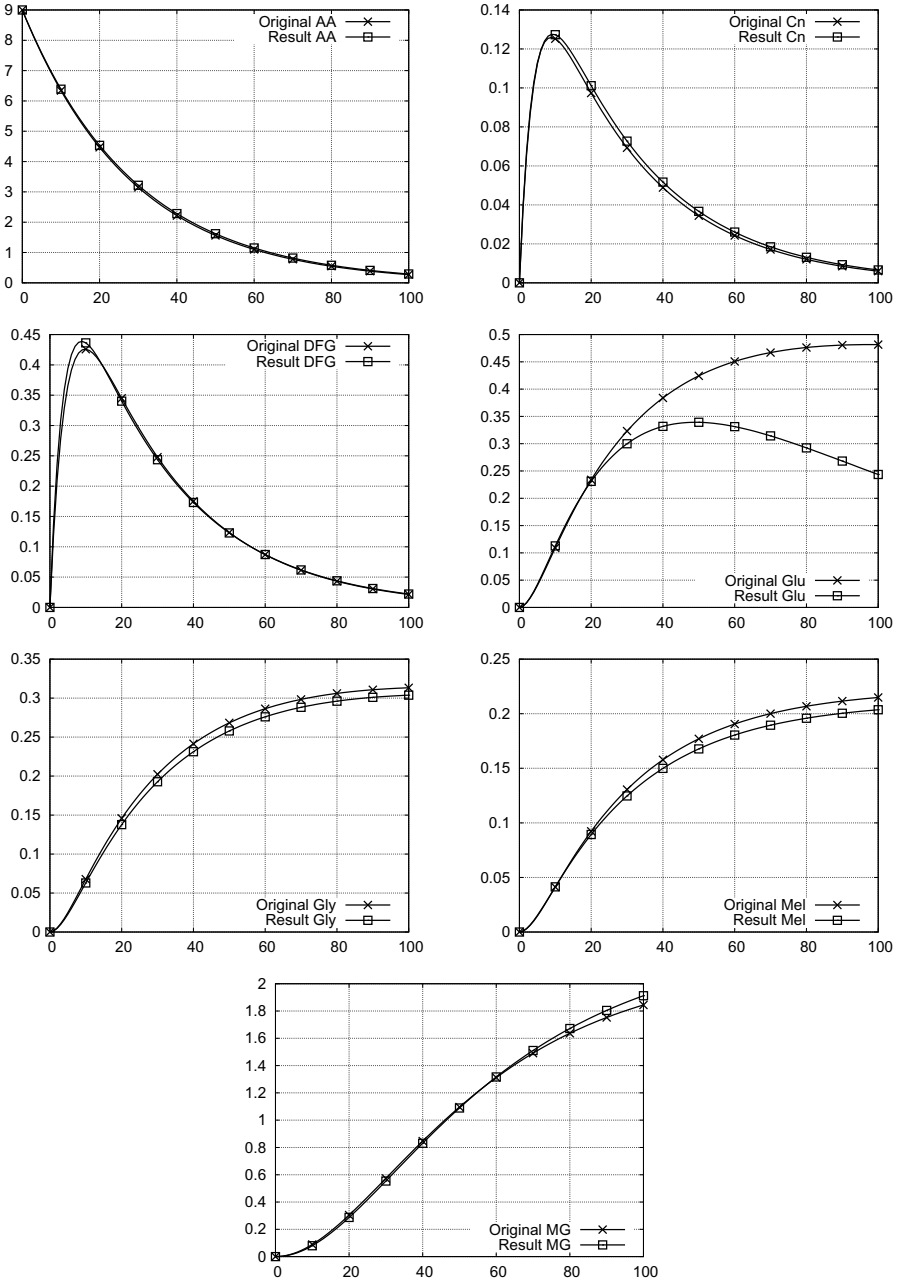

Fig. 5. ODEs of some of the reagents involved in the DFG degradation pathway

of observations of the modeled phenomenon. Several numerical examples have been provided to illustrate the computational characteristics of the method.

References

[1] Boys, R. J., D. J. Wilkinson and T. B. L. Kirkwood, *Bayesian inference for a discretely observed stochastic kinetic model,*, Statistics and Computing **18** (2008), pp. 125–135.

[2] Burrows, R., G. Warnes and R. Choudary Hanumara, *Statistical modeling of biochemical pathways*, Technical Report 06/11, Dept. of Biostatistics and Computational Biology, University of Rochester (2006).

[3] Dempster, A. P., N. M. Laird and D. B. Rubin, *Maximum likelihood from incomplete data via the em algorithm*, JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B **39** (1977), pp. 1–38.

[4] Gadkar, K. G., R. Gunawan and F. J. Doyle 3rd, *Iterative approach to model identification of biological networks*, BMC Bioinformatics **6** (2005).

[5] Gillespie, D. T., *Exact stochastic simulation of coupled chemical reactions*, J. Phys. Chem. **81** (1977), pp. 2340–2361.

[6] Golding, I., J. Paulsson, S. M. Zawilski and E. C. Cox, *Real-time kinetics of gene activity in individual bacteria*, Cell **123** (2005), pp. 1025–1036.

[7] Golightly, A. and D. Wilkinson, *Bayesian inference for stochastic kinetic models using a diffusion approximation*, Biometrics **61** (2005), pp. 781–788.

[8] Kurtz, T. G., *Solutions of ordinary differential equations as limits of pure jump Markov processes*, Journal of Applied Probability **1** (1970), pp. 49–58.

[9] Martins, S. I., A. T. Martinus and M. A. V. Boekel, *Kinetic modelling of amadori n-(1-deoxy–fructos-1-yl)-glycine degradation pathways. part ii–kinetic analysis*, Carbohydrate Research **338** (2003), pp. 1665–1678.

[10] Moles, C., P. Mendes and J. Banga, *Parameter estimation in biochemical pathways: a comparison of global optimization methods*, Genome Res. **13** (2003), pp. 2467–2474.

[11] Reinker, S., R. Altman and J. Timmer, *Parameter estimation in stochastic chemical reactions*, IEEE Proceedings Systems Biology **153** (2006), pp. 168–178.

[12] Sugimoto, M., S. Kikuchi and M. Tomita, *Reverse engineering of biochemical equations from time-course data by means of genetic programming*, Biosystems **80** (2005), pp. 155–164.

[13] Tian, T., S. Xu, J. Gao and K. Burrage, *Simulated maximum likelihood method for estimating kinetic rates in gene expression*, Bioinformatics **23** (2007), pp. 84–91.

[14] Tribastone, M., "Scalable Analysis of Stochastic Process Algebra Models," Ph.D. thesis, School of Informatics, University of Edinburgh (2010).

[15] Wei, G. C. G. and M. A. Tanner, *A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms*, Journal of the American Statistical Association **85** (1990), pp. 699–704.